데이터시각화 기말과제물

**1. 런던 콜레라지도(교재 p.4)를 조사하고 데이터 시각화의 필요성을 설명하시오.(7점)**

**2. 아름답고 의미 있는 데이터 시각화 사례를 발굴하고 그 이유를 정리하시오.(7점)**

**3. R 패키지 “vcd”에 내장된 “Arthritis” 데이터셋은 류마티스 관절염 환자를 대상으로 한 임상시험 결과 데이터이다. 각 행은 각 환자를 나타내며, 변수 Treatment는 그룹 (Treated = 새로운 치료제를 투약한 그룹, Placebo = 위약을 받은 그룹)을 나타낸다. 변수 Sex는 성별을, Improved는 치료 결과(None = 차도 없음, Some = 약간 좋아짐, Marked = 매우 좋아짐)를 나타낸다. 새로운 치료제 투약 여부가 치료 결과와 연관이 있는지, 성별과 치료 결과 간에 연관이 있는지를 데이터 시각화를 통해서 탐구하시오. (18점)**

library(vcd)

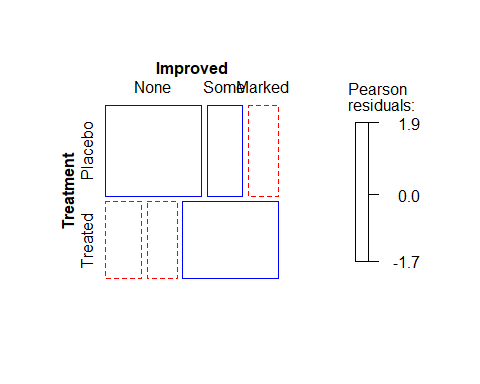
## Warning: 패키지 'vcd'는 R 버전 4.1.3에서 작성되었습니다

## 필요한 패키지를 로딩중입니다: grid

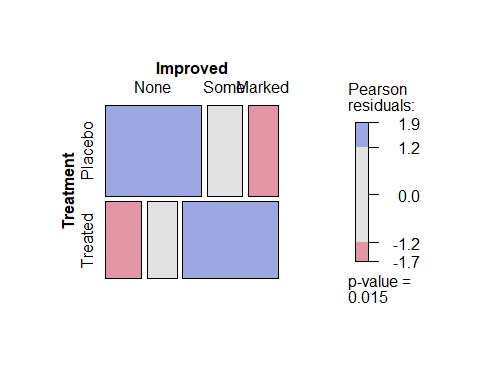
data("Arthritis")  
art <- xtabs(~ Treatment + Improved, data = Arthritis, subset = Sex == "Female")  
art

## Improved  
## Treatment None Some Marked  
## Placebo 19 7 6  
## Treated 6 5 16

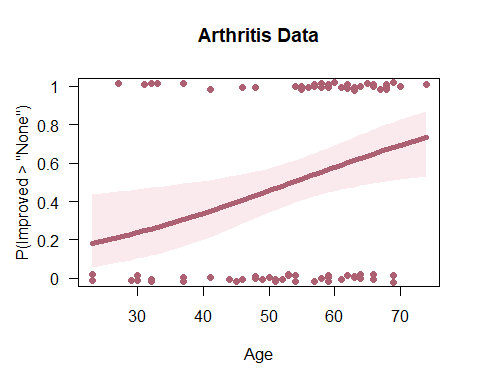
mosaic(art, gp = shading\_Friendly)



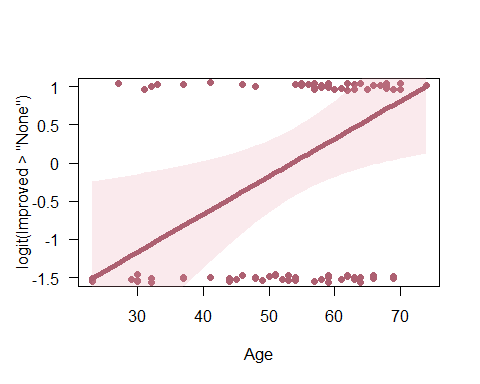
mosaic(art, gp = shading\_max)



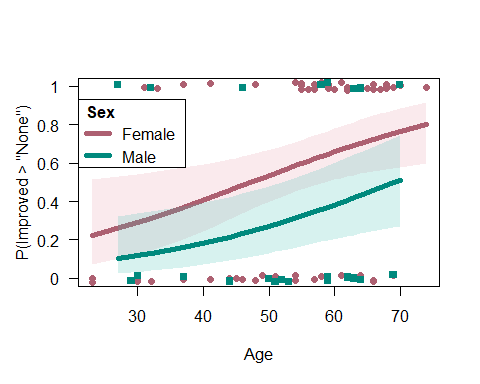
library(vcd)  
## Simple model with no conditioning variables  
art.mod0 <- glm(Improved > "None" ~ Age, data = Arthritis, family = binomial)  
binreg\_plot(art.mod0, "Arthritis Data")



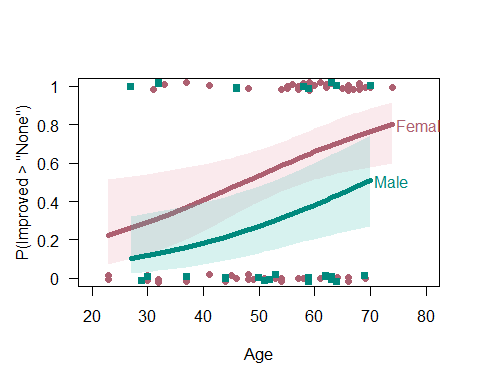
binreg\_plot(art.mod0, type = "link") ## logit scale



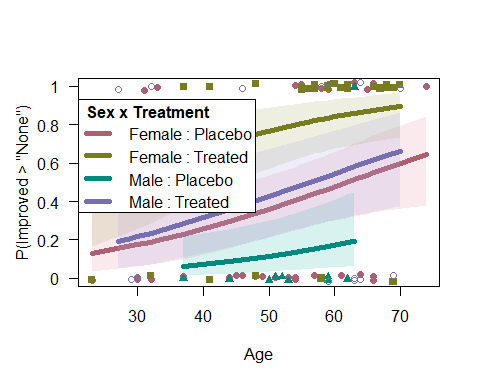
## one conditioning factor  
art.mod1 <- update(art.mod0, . ~ . + Sex)  
binreg\_plot(art.mod1)



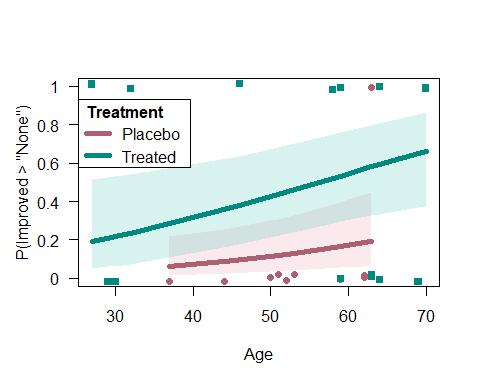
binreg\_plot(art.mod1, legend = FALSE, labels = TRUE, xlim = c(20, 80))



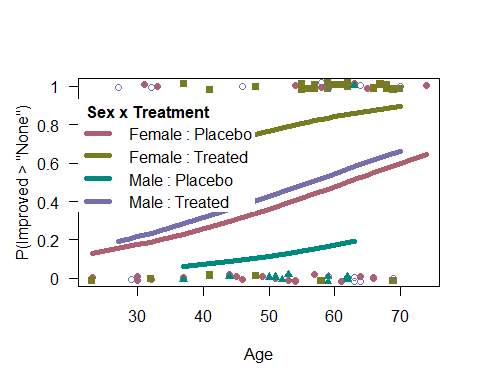
## two conditioning factors  
art.mod2 <- update(art.mod1, . ~ . + Treatment)  
binreg\_plot(art.mod2)



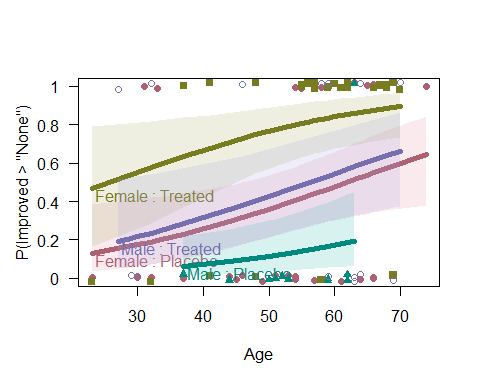
binreg\_plot(art.mod2, subset = Sex == "Male") ## subsetting



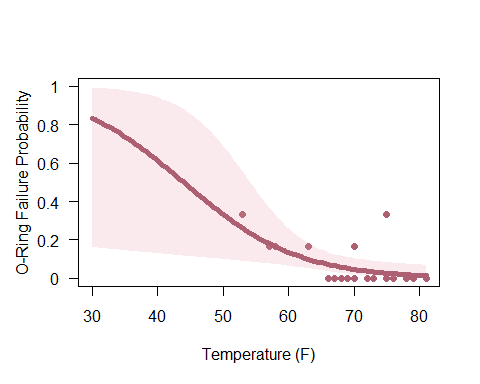
## some tweaking  
binreg\_plot(art.mod2, gp\_legend\_frame = gpar(col = NA, fill = "white"), col\_bands = NA)



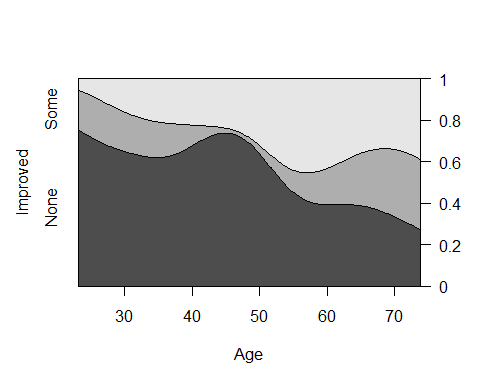
binreg\_plot(art.mod2, legend = FALSE, labels = TRUE,  
labels\_pos = "left", labels\_just = c("left", "top"))



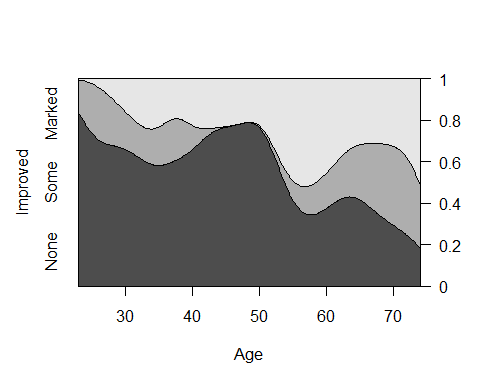
## model with grouped response data  
shuttle.mod <- glm(cbind(nFailures, 6 - nFailures) ~ Temperature,  
data = SpaceShuttle, na.action = na.exclude, family = binomial)  
binreg\_plot(shuttle.mod, xlim = c(30, 81), pred\_range = "xlim",  
ylab = "O-Ring Failure Probability", xlab = "Temperature (F)")



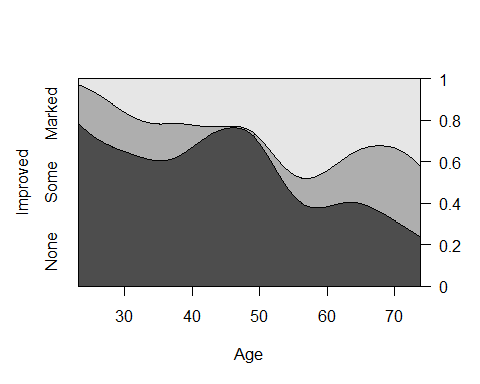
## Arthritis data  
data("Arthritis")  
cd\_plot(Improved ~ Age, data = Arthritis)



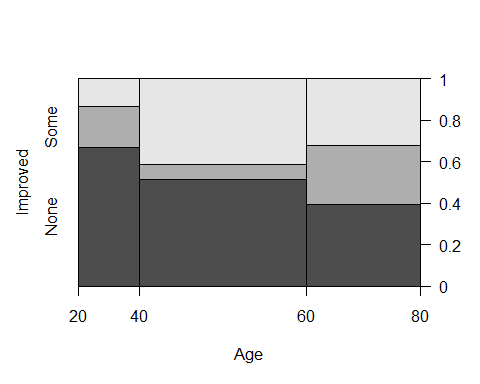
cd\_plot(Improved ~ Age, data = Arthritis, bw = 3)



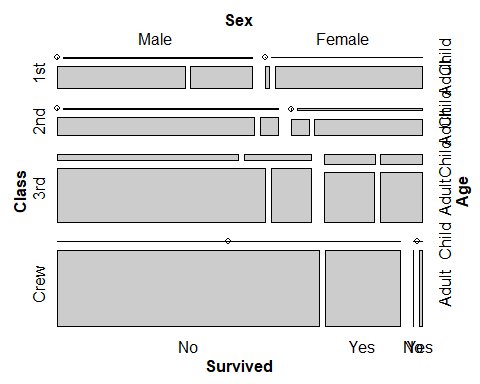
cd\_plot(Improved ~ Age, data = Arthritis, bw = "SJ")



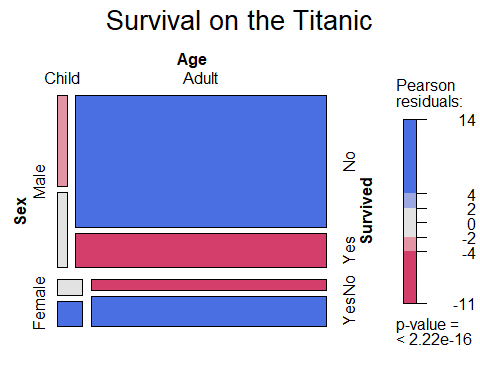
## compare with spinogram  
spine(Improved ~ Age, data = Arthritis, breaks = 3)



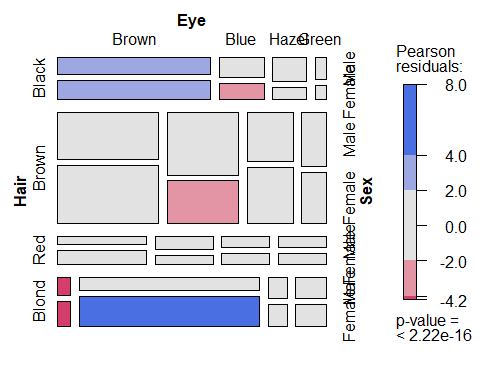
library(MASS)  
data("Titanic")  
mosaic(Titanic)



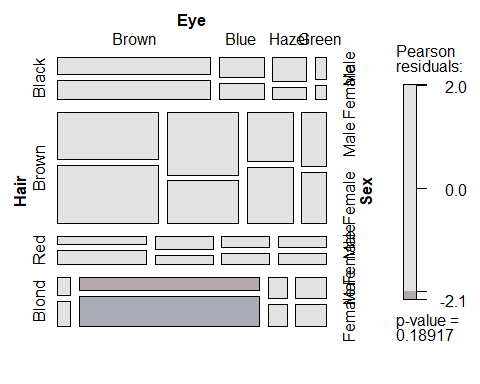
## Formula interface for tabulated data plus shading and legend:  
mosaic(~ Sex + Age + Survived, data = Titanic,  
main = "Survival on the Titanic", shade = TRUE, legend = TRUE)



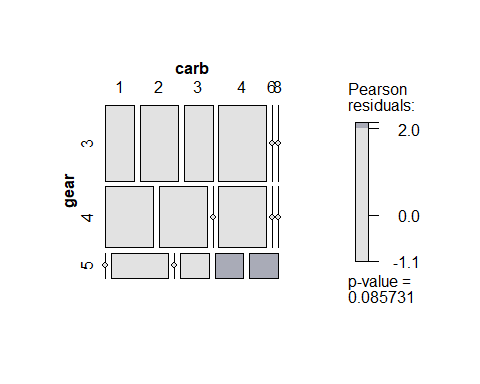
data("HairEyeColor")  
mosaic(HairEyeColor, shade = TRUE)



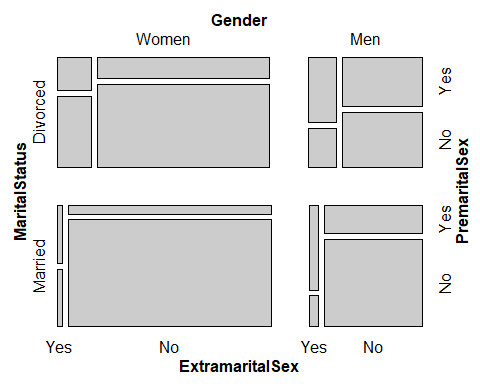
## Independence model of hair and eye color and sex. Indicates that  
## there are significantly more blue eyed blond females than expected  
## in the case of independence (and too few brown eyed blond females).  
mosaic(HairEyeColor, shade = TRUE, expected = list(c(1,2), 3))



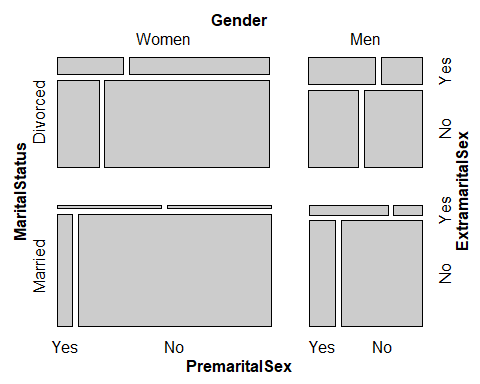
## Model of joint independence of sex from hair and eye color. Males  
## are underrepresented among people with brown hair and eyes, and are  
## overrepresented among people with brown hair and blue eyes, but not  
## "significantly".  
## Formula interface for raw data: visualize crosstabulation of numbers  
## of gears and carburettors in Motor Trend car data.  
data("mtcars")  
mosaic(~ gear + carb, data = mtcars, shade = TRUE)



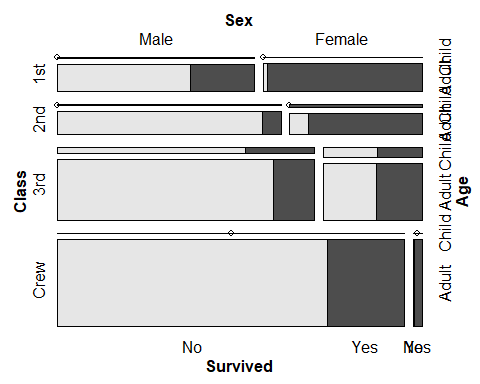
data("PreSex")  
mosaic(PreSex, condvars = c(1,4))



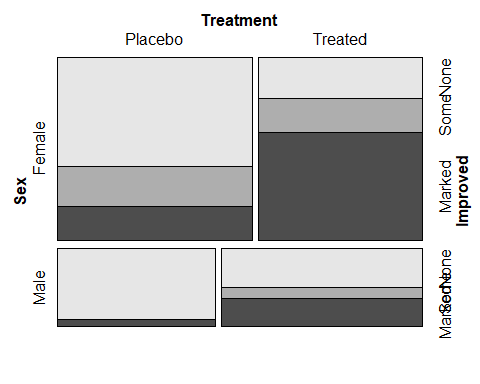
mosaic(~ ExtramaritalSex + PremaritalSex | MaritalStatus + Gender,  
data = PreSex)



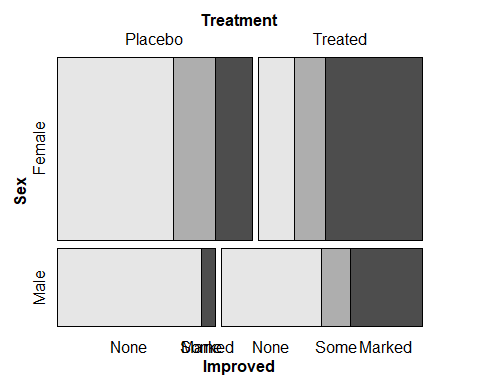
## Highlighting:  
mosaic(Survived ~ ., data = Titanic)



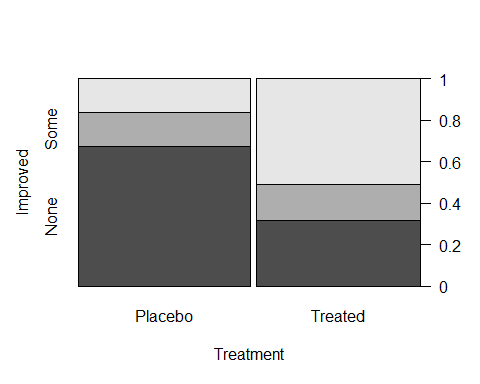
data("Arthritis")  
mosaic(Improved ~ Treatment | Sex, data = Arthritis, zero\_size = 0)



mosaic(Improved ~ Treatment | Sex, data = Arthritis, zero\_size = 0,  
highlighting\_direction = "right")

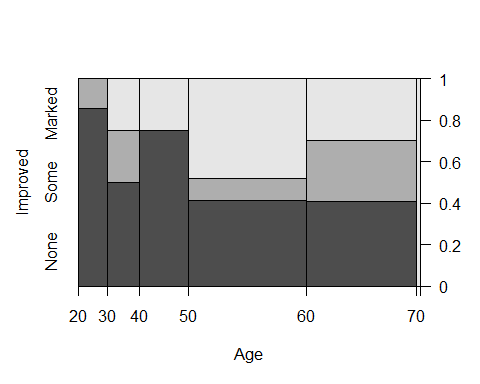


## Arthritis data (dependence on a categorical variable)  
data("Arthritis")  
(spine(Improved ~ Treatment, data = Arthritis))



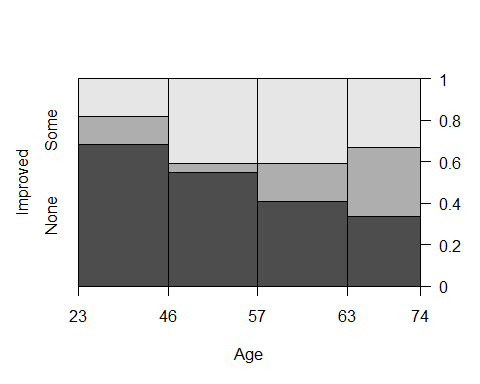
## Improved  
## Treatment None Some Marked  
## Placebo 29 7 7  
## Treated 13 7 21

## Arthritis data (dependence on a numerical variable)  
(spine(Improved ~ Age, data = Arthritis, breaks = 5))



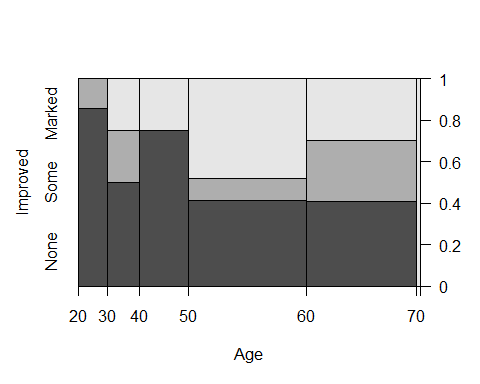
## Improved  
## Age None Some Marked  
## [20,30] 6 1 0  
## (30,40] 4 2 2  
## (40,50] 9 0 3  
## (50,60] 12 3 14  
## (60,70] 11 8 8  
## (70,80] 0 0 1

(spine(Improved ~ Age, data = Arthritis, breaks = quantile(Arthritis$Age)))



## Improved  
## Age None Some Marked  
## [23,46] 15 3 4  
## (46,57] 12 1 9  
## (57,63] 9 4 9  
## (63,74] 6 6 6

(spine(Improved ~ Age, data = Arthritis, breaks = "Scott"))



## Improved  
## Age None Some Marked  
## [20,30] 6 1 0  
## (30,40] 4 2 2  
## (40,50] 9 0 3  
## (50,60] 12 3 14  
## (60,70] 11 8 8  
## (70,80] 0 0 1

**4. R에 내장된 “airquality” 데이터셋은 1973년 5월부터 9월까지 뉴욕의 대기질에 관한 데이터셋이다. 변수 Ozone은 대기 중 오존의 양, Solar.R은 태양방사선의 양, Wind는 풍속, Temp는 기온을 나타낸다. 이 네가지 변수(Ozone, Solar.R, Wind, Temp)에 대한 산점도 행렬을 그리고, 이 산점도 행렬에서 알 수 있는 변수들 간의 관계에 대하여 서술하시오. (18점)**